Voice-Image Cross-Modal Human Fatigue Detection Based on CNN-ELM Hybrid Model

Shuxi Chen[1], Yiyang Sun[1], Jianlin Qiu[2], Haifei Zhang[1], Hao Chen[3]*

[1] School of Computer and Information Engineering, Nantong Institute of Technology, China
[2] School of Information Science and Technology, Nantong University, China
[3] School of Electrical Engineering and Computer Science, Royal Institute of Technology (KTH), Sweden
chenshq@ntit.edu.cn, sunyy@ntit.edu.cn, qiu.jl@ntu.edu.cn, 20160063@ntit.edu.cn, haoch@kth.se

Abstract

The deepening of human fatigue will lead to the reduction of exercise ability and work efficiency, the increase of errors and accidents, and even the occurrence of organic diseases. Obviously, it is significant to understand the impact of human fatigue on the health, safe production and safe work of different people. At present, fatigue detection is mostly carried out through EEG and EMG signals. These methods usually have the disadvantages of contact and non-realtime.

In response to the aforementioned issues in the process of human fatigue detection, this article effectively applies the visual image analysis method of spectrograms to human fatigue detection and proposes a cross-modal human fatigue detection method based on speech spectral image recognition. First, Mel spectrograms of speech segments in the corpus are extracted, and a fatigue spectrogram data set is established. A deep learning model is established through convolutional neural network (CNN) and extreme learning machine (ELM) for spectral image recognition and fatigue detection. CNN is used to extract features from the input image. The feature mapping will eventually be encoded into a one-dimensional vector and sent to ELM for classification. The experimental results indicate that the speech spectrum image features extracted by this method have better fatigue characterization ability than traditional acoustic features.

Keywords: Fatigue, Mel spectrogram, Cross-modal, CNN, ELM

1 Introduction

Nowadays, with the increase of work pressure and the acceleration of the pace of life, people often have muscle soreness, drowsiness and other fatigue phenomena, which seriously affects people's physical and mental health and disturbs pace of life and work state. Research shows that after fatigue, people's cardiovascular and neural functions will change, and their will will be weaken, their attention will not be concentrated, and their information processing will be slow and chaotic, resulting in low work efficiency and even mistakes, which is difficult to ensure safety and efficiency. Therefore, the detection of fatigue is particularly important by recognizing the harm of continuing to work after fatigue [1].

We can divide fatigue testing methods into two ways: subjective and objective testing. Subjective detection method refers to the assessment of fatigue based on subjective perception, Subjective questionnaires, Stanford Sleep Scale, and Sleep Habits Questionnaire are commonly used for fatigue assessment. Objective methods can be divided into contact method and non-contact method. Contact method is used to detect fatigue through EEG signals and EMG signals [2-3]. The non-contact method is mainly used to detect the fatigue degree by analyzing the facial expression and posture of people. The artificial intelligence and big data technology can be used to accurately predict the fatigue status of the human body and classify the corresponding fatigue level, so that the user can obtain and perceive the fatigue information. Although the methods mentioned above can measure the fatigue level, there is a certain degree of operability in practical application. Subjective detection method has a low accuracy rate due to excessive subjective factors. Objective methods cannot achieve real-time detection, and has certain invasiveness, which will make users feel resistant.

Considering the limitations of the fatigue detection methods mentioned above, this paper adopts a cross-modal human fatigue detection method which is based on spectrogram features. Speech contains rich content and emotional information, so we often transmit information through speech. Fatigue detection by voice can effectively avoid the shortcomings of the above fatigue detection methods. Therefore, the academic community began to shift the research focus to fatigue detection based on speech analysis. In the speech detection method of human fatigue state, the extraction of fatigue related speech features is the most critical link, which will directly determine the detection effect [4]. The fatigue related speech features commonly used in the existing research are mainly concentrated in the traditional acoustic features, including sound quality, prosody, frequency domain and nonlinear features [5]. However, these features are often limited to a single time-domain or frequency-domain analysis, ignoring the correlation between human fatigue

*Corresponding Author: Shuxi Chen; Email: chenshq@ntit.edu.cn
DOI: https://doi.org/10.70003/160792642025092605003

and the common influence of speech signals in time and frequency domain, resulting in the analysis angle not comprehensive and detailed, and the detection effect needs to be improved [6]. In recent years, the two-dimensional visual image analysis method for speech spectrum [7-8] has provided a new idea for human fatigue detection.

We have summarized the contributions of the study as follows:

1. We propose a cross-modal fatigue detection method that converts fatigue detection through speech into image recognition of speech spectrograms, thereby achieving the purpose of fatigue detection.

2. We propose a hybrid model based on CNN-ELM for fatigue spectrogram image recognition.

3. The corpus and extracted spectrogram image dataset associated with the project have been contributed to GitHub to facilitate comparative research by other researchers. (https://github.com/csx0709/corpus; https://github.com/csx0709/spectrogram)

We have structured our paper in the following way. Section 2 gives a brief overview of the principle of detecting fatigue level through speech spectrum, and a brief statement of the technical basis of the study on neural networks. Section 3 briefly describes the voice-image fatigue detection process based on CNN-ELM hybrid model. The method of extracting spectrograms, and image recognition method based on a mixture of CNN and ELM also introduced in this section. Section 4 describes the motivation for our study and entire experimental procedure. Section 5 provides an overview of future research directions and offers research insights.

2 Background

In this section we would like to provide the principle of detecting fatigue level through speech spectrum. Then, we give a brief statement of the technical basis of the study on neural networks.

2.1 Influence Mechanism of Human Fatigue on Spectrogram

In the field of speech recognition, it is different from the "human short-term state detection" of speech emotion recognition (emotional state is mainly affected by psychological factors and may change dramatically in a short time) and the "human long-term state detection" of biometric recognition such as gender, age and identity (such state will not change greatly in a long period of time). Human fatigue detection belongs to the "human medium-term state detection" (also including drunk state and disease state). Such medium-term state often has a negative impact on the human body and its vocal system for tens of minutes or even days, and it is difficult or impossible to subjectively control the change of state. The influence of human fatigue on vocal organs and voice signals is mainly reflected in: When the human body is tired, the sound pressure energy decreases due to the slow breathing; Vocal cord relaxation leads to the decrease of gene frequency; Brain vitality and speech planning ability decreased, resulting in reduced speech intelligibility, abnormal prosody and slower speech speed; The relaxation of vocal tract, throat and facial muscles will make the formant frequency in the spectrum decrease as a whole, and the bandwidth becomes wider; the changes in the temperature and viscoelasticity of the vocal tract wall will increase the friction between the vocal tract wall and the speech air flow, thereby further reducing the position of the formant and correspondingly increasing the bandwidth. This phenomenon is particularly obvious in the low-frequency part of 200-3000 Hz in the voiced section [9]; The most important thing is that it is very likely to have a wheezing period or a silent period in the fatigue state. This fatigue effect can be directly shown in the spectrogram.

2.2 Neural Networks

With the progress of science and technology, a strong idea of simulating the human brain system has gradually emerged. On this idea, multiple scholars have persistently explored and laid the foundation for the development of neural networks. In 1943, McCulloch and Pitts attempted to understand how the brain produces complex patterns through interconnected neurons, and MCP model was proposed [10]. In 1965, Ivakhnenko and Lapa proposed a supervised deep feedforward multilayer perceptron algorithm [11]. The term deep learning was first used and described in 2006 [12]. Artificial neural networks were regarded as having the ability to learn the essential characteristics from raw data. This is one of the most significant breakthroughs in the field of deep learning. We learned this ability step by step through a multilayer neural network [13]. Different from feature engineering in traditional machine learning, this learning method was called end-to-end learning. In 2010, Deep Convolutional Neural Networks (CNN) were introduced for ImageNet classification in the Challenge. The results on the test data show that CNNs clearly outperform other methods [14].

CNN is often used to process data with grid structure and it can easily handle high latitude data. CNN was originally proposed to solve the problem of image recognition, commonly used to handle problems in the field of computer vision, such as image classification, object detection, and image segmentation. It aims to capture local features of input data through convolution operations in net-worked hierarchical structure, and to decrease the size of the feature map through pooling operations. CNN can automatically learn and extract various features in images, such as edges, textures, shapes, etc [15-17]. CNN's success is largely attributed to its ability to automatically learn feature representation through the Backpropagation, thus avoiding the tedious process of manually extracting features [18].

At present, using a hybrid model built jointly with CNN and SVM is the most common combination method. Matsugu and Ebert validated that the hybrid model of CNN and SVM has more advantages than the CNN model. Due to the complexity of the design of SVM, the SVM algorithm achieves probability prediction through the expensive cross validation method. Therefore, it is of great significance to find classifiers with fewer parameters,

good classification performance, and strong generalization ability to handle image classification tasks.

3 Our Method

In this section, we briefly describe the voice-image fatigue detection process based on CNN-ELM hybrid model. We introduce the method of extracting spectrograms, and image recognition method based on a mixture of CNN and ELM.

3.1 Fatigue Detection Process Based on Spectrogram Analysis

Speech spectrum image, also referred to as spectrogram, can visually express the joint distribution of speech signal in time and frequency domain. It can also analyze and extract relevant image features from it can break through the singularity of traditional acoustic features, acquire more speech information that cannot be characterized by traditional acoustic features from the perspective of joint analysis in time and frequency domain, and also make some very meaningful achievements in many speech recognition fields.

Therefore, fatigue detection based on speech spectrum and image features is feasible. It is mainly used to carry out cross modal research on acoustic signals, extract the spectrum image of speech, construct the Mel spectrum image data set of speech, and use the deep learning method to learn the image to detect fatigue. The main flow chart is shown in Figure 1.

Figure 1. Flow chart of speech fatigue detection based on spectrogram analysis

3.2 Generation of Speech Spectrogram

The speech signal $s(t)$ continuously sampled in the time domain is processed by framing, windowing and short-time discrete Fourier transform, including:

$$X(n,f)=\frac{1}{N}\sum_{t=0}^{N-1}h(t)s(t,n)e^{-j\frac{2\pi}{N}f\cdot t} \quad (1)$$

Where: t refers to time domain sampling points; n refers to speech frames; f refers to frequency points; $s(t,n)$ represents the speech signal of the nth frame; $h(t)$ is the window function; N is the window length. Then, the logarithmic energy spectrum of the speech signal can be calculated as follows:

$$F(n,f)=20\lg\left[X(n,f)\right] \quad (2)$$

The spectrogram $f(n,f)$ thus obtained is a two-dimensional image that can intuitively represent the change of speech spectrum with time. The horizontal axis n corresponds to time and the vertical axis f corresponds to frequency. The value of each coordinate point corresponds to the energy of a frequency component at a certain time. Accordingly, the change of resonance frequency of sound with time will show different visual image texture changes in spectrogram [9].

The spectrogram is often very large. The spectrogram is often transformed into Mel spectrum by Mel scale filter banks, so that we can acquire sound features at appropriate size. Sound level heard by the human ear does not have a linear relationship with the actual frequency. Mel frequency is more in line with the auditory characteristics of the human ear. Linear distribution below 1000Hz and logarithmic increase above 1000Hz. The transformation method between Mel frequency and linear frequency is:

$$Mel(f)=2595*lg(1+f/700) \quad (3)$$

The mapping relationship between the linear frequency and the Mel frequency described in this formula is shown in Figure 2.

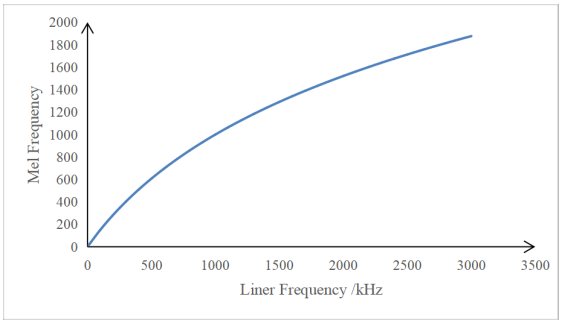

Figure 2. Mapping relationship between linear and Mel frequency

The acquisition process of Mel spectrum diagram is shown in Figure 3.

Figure 3. Flow chart of Mel spectrum acquisition

3.3 CNN-ELM Hybrid Model

The principle of hybrid model CNN-ELM is to use ELM to replace the trained output layer of CNN for classification, and its schematic diagram is shown in Figure 4. The hybrid model mainly includes two stages, feature extraction and classification. Firstly, CNN parameters are adjusted in the training process until the network structure reaches the desired state. Then ELM is called to receive the feature map encoded into a one-dimensional vector, and ELM gives the classification result.

3.3.1 CNN

The main components of CNN include the convolutional layer, activation function, pooling layer and fully connected layer. Among them, convolutional layer is the foremost part of CNN. It extracts features by

applying a series of convolutional kernels (which we also call filters) on the input data. The activation function is usually inserted behind the convolution layer to introduce nonlinear properties. Th pooling layer can reduce the size of the feature map, so that we can extract more prominent features. Finally, the fully connected layer connects the output of pooling layer to one or more neurons to produce the final prediction result [19-20]. The structure of CNN is shown in Figure 5.

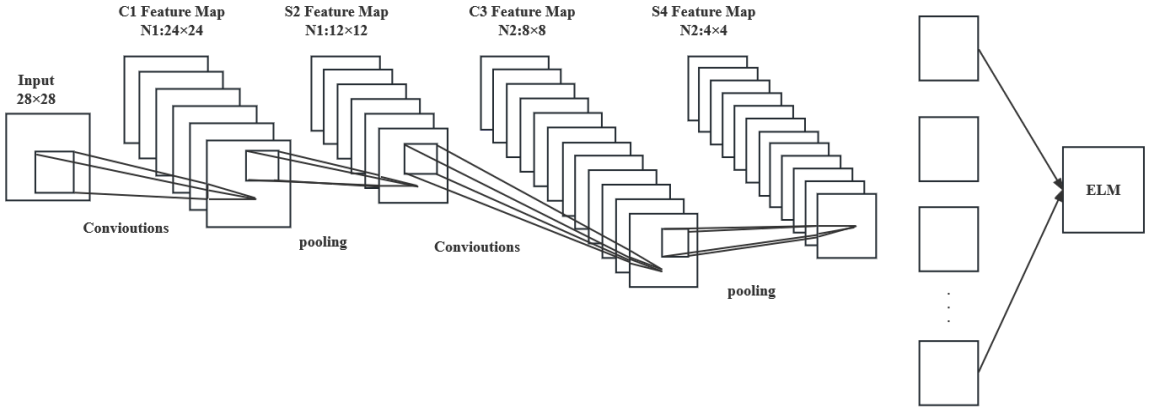

Figure 4. Schematic diagram of CNN-ELM hybrid model

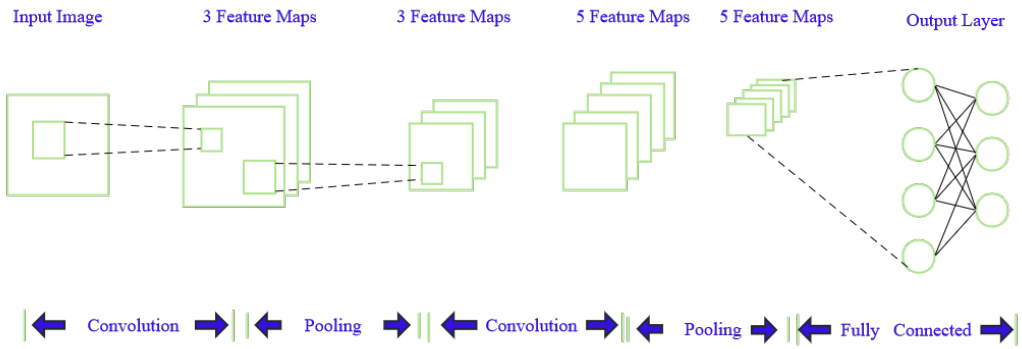

Figure 5. Structure of CNN

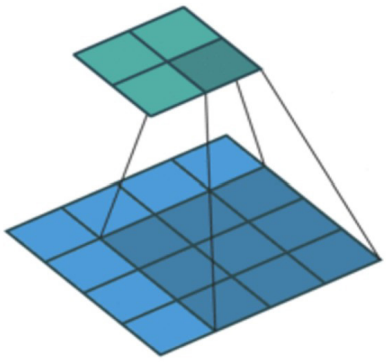

Figure 6. Schematic diagram of convolution operation

(1) Convolution layer

Convolutional layer is one of the key components in CNN, whose main purpose is to extract local patterns. Convolutional operation can enhance the original signal and reduce noise. In addition, in each feature map, the weights of each filtering factor are shared, which not only reduces the free parameters of the network, but also reduces the complexity of the relevant layers. The output of convolution operation includes multiple feature maps, and each neuron in the entire feature map connects to the local region of the previous layer. Convolutional layer includes: convolutional kernels, feature maps, stride and padding. The schematic diagram of convolution operation is shown in Figure 6.

The convolution operation can be expressed by the equation (4).

$$F_l^k = \left(I_{x,y} * K_l^k\right) \tag{4}$$

Where $I_{x,y}$ refers to an input image. X and Y refer to positions. K_l^k is the lth convolution kernel (Layer K).

(2) Pooling layer

We can consider the pooling layer as a fuzzy filter. It is used to re-extract features from the convolutional layer. Equation (5) represents the operation of the pooling layer [21].

$$Z_l = f_p\left(F_{x,y}^l\right) \tag{5}$$

By utilizing the principle of local correlation, pooling operation not only eliminates non maximum values and reduces the number of parameters in the previous layer, but also improves the network's anti distortion ability. Like the convolution layer, the output terminals of a limited number of neurons in the Receptive field in the previous layer are connected to each neuron in the pooling layer. The goal of pooling is to perform secondary sampling on the input image, reduce computational load, and avoid overfitting.

The most commonly used pooling methods include average pooling and maximum pooling. Figure 7 shows the operation diagram of maximum pooling, which reduces the feature map with an input of 4×4 and an output of 2×2. Boureau built multiple sets of experiments, summarized experimental data, and conducted a detailed analysis of their performance, attempting to find their characteristics. Scherer made in-depth research and exploration, and found that maximum pooling can accelerate Rate of convergence, have better feature invariance, and improve generalization performance.

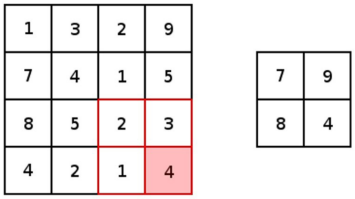

Figure 7. Schematic diagram of maximum pooling

(3) Activation function

Regardless of the input layer, hidden layer, or output layer, all layers have nodes, and each node has a weight that is considered when processing information from the previous layer to the next layer. The activation function can often play a decision-making role, and has an important auxiliary role in the learning of those complex patterns. Selecting a reasonable and appropriate activation function has an important accelerating effect on the learning process. If an activation function is used, the output signal will no longer be a simple linear function. Although linear equations are simple and easy to solve, they lack the ability to learn and identify complex mappings from data due to their limited complexity. In practical applications, it is not only hoped that the model can learn and calculate various linear functions, but also hopes to understand complex, high-dimensional, and nonlinear datasets. Equation (6) defines the activation function. Activation function diagram is shown in Figure 8.

$$T_l^k = f_A\left(F_l^k\right) \tag{6}$$

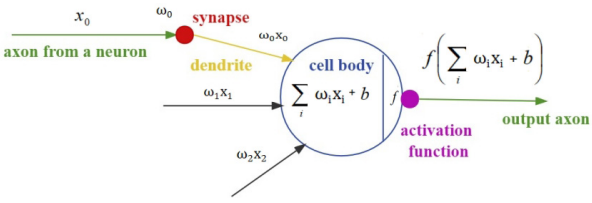

Figure 8. Diagram of activation function

(4) Dropout

Dropout introduces regularization (within the network) into it, and with a certain probability, some units or connections are skipped in a random manner, so as to improve its generalization. In the whole neural network architecture, when learning various connections corresponding to a certain nonlinear relationship, sometimes there will be mutual adaptation, which will cause overfitting [22]. In addition, it should also be noted that the random discarding of some connections or units will result in several network architectures. Among them, one network with relatively small weight and certain representativeness can be selected, and the selected architecture is regarded as the approximation of all networks [23].

(5) Full connection layer

In general, the full connection layer is mainly used for the classification task at the end of the network, which is different from convolution and pooling. Among them, the biggest difference is that it is a global operation, which can achieve the input of one layer and perform global analysis on the output of the previous layer. With the help of the selected features, nonlinear combination is implemented, which can be used to classify the data [24-25]. In the fully connected layer, features are encoded into one-dimensional vectors, which are then classified using a trainable classifier. The fully connected layer is shown in Figure 9.

(6) Training process of CNN

The purpose of training CNN is to adjust the parameters of the entire system, namely the weights and deviations of the convolutional kernel, and use the fine-tuned CNN to classify and predict unknown input image data. The training flow chart is shown in Figure 10.

CNN training can be divided into two processes. The first process is the propagation phase from lower to higher levels, which is forward propagation. The second stage is when the results obtained from the current propagation do not match the expectations, and the error propagates from higher layers to lower layers, that is, back propagation [26].

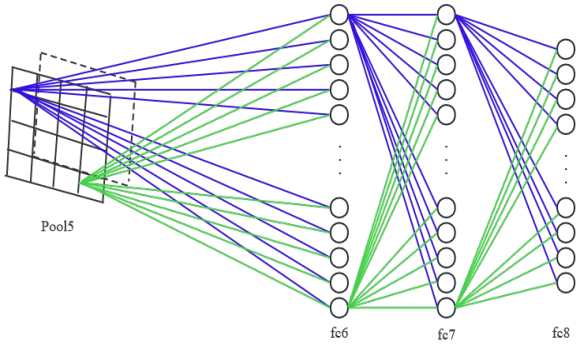

Figure 9. Schematic diagram of fully connection layer

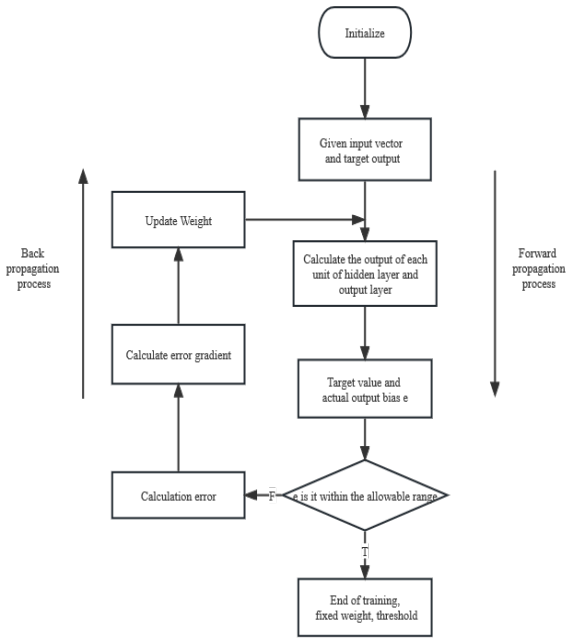

Figure 10. Training process

3.3.2 ELM

As a fast learning algorithm, ELM randomly initializes input layer weights and hidden layer biases during the training process, and finally calculates the optimal solution through generalized inverse matrix theory. The network structure of ELM is shown in Figure 11.

Assuming there are N training samples, $x_i = [x_{i1}, x_{i2}, ..., x_{in}]^T \in R^n$, $t_i = [t_{i1}, t_{i2}, ..., t_{im}]^T \in R^m$, x_i refers to input vectors. t_i refers to output vectors. A single hidden layer neural network with L hidden layer nodes can be represented as:

$$\sum_{i=1}^{L} \beta_i g(\omega_i \cdot x_j + b_j) = o_j, j = 1 ... N \quad (7)$$

Among them, g(x) is activate function. $\omega_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{in}]^T$ is input weights. β_i is output weights. b_i is the bias of the ith hidden layer neuron. The aim of single hidden layer neural networks is o_j infinitely approaching t_j. So there is:

$$\sum_{i=1}^{L} \beta_i g(\omega_i \cdot x_j + b_j) = t_j, j = 1 ... N \quad (8)$$

The matrix is represented as:

$$H\beta = \mathrm{T} \quad (9)$$

$$[h_{ij}] = \begin{bmatrix} g(\omega_1 \cdot x_1 + b_1) & \cdots & g(\omega_L \cdot x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(\omega_1 \cdot x_N + b_1) & \cdots & g(\omega_L \cdot x_N + b_L) \end{bmatrix} \quad (10)$$

$$\beta = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{L1} & \cdots & \beta_{Lm} \end{bmatrix} \quad (11)$$

$$\mathrm{T} = \begin{bmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{N1} & \cdots & t_{Nm} \end{bmatrix} \quad (12)$$

H is the output of the hidden layer node. T is the expected output. Equation (9) represents a linear system whose the only optimal solution can be calculated using the least squares method.

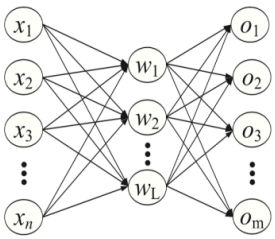

Figure 11. Diagram of ELM

4 Implementation

In this section, we first briefly describe the motivation for our study. Based on this research question, we then describe our entire experimental procedure from various perspectives, including experimental design.

4.1 Research Questions

The fatigue detection performance of CNN-ELM is our primary concern when conducting research, so we must first address the question of whether our method is optimal. Research questions (RQs) are designed for our study:

RQ1: Whether better classification performance than CNN can be achieved using our designed method?

The motivation of RQ1 is that we compare our proposed approach with CNN techniques [27]. We also compare our method with some popular object detection algorithm, including: Mask R-CNN [28], MR R-CNN [29] and Faster R-CNN [30].

RQ2: Whether better classification and speed performance than other classifier can be achieved using our designed method?

The motivation of RQ2 is that we compare our proposed approach with SVM.

RQ3: Whether our proposed approach solve the problem of overfitting?

The motivation of RQ3 is that we adopt cross-validation when we conduct experiments, so that we can avoid overfitting and data leakage during the training process.

4.2 Corpus

Considering the influence of the quality of the original speech signal on the recognition performance in our system, and the lack of a dedicated fatigue detection corpus within the academic team, the preliminary work of this paper includes the establishment of the Soochow University Speech Processing–Sport Fatigue Detection (SUSP-SFD) Corpus [31], which is available at: https://github.com/csx0709/corpus. During the corpus construction process, we recruited 30 native Chinese speakers for corpus recording, which made our corpus rich and varied. Our corpus recording condition is dual channel, with a sampling frequency of 48kHz. At the same time, we use heart rate as an auxiliary judgment basis for fatigue level to evaluate the reliability and effectiveness of the corpus.

4.3 Spectrogram Dataset

In this paper, the third-party library librosa of Python speech signal processing is used to extract spectrogram. First, using librosa.filters.mel() to create a filter bank matrix to merge FFT into Mel frequency. Then, according to the audio time series and sampling rate, the amplitude spectrum of the audio is calculated, and then through mel_f.dot() function maps it to mel scale. The whole process has been encapsulated in the librosa.feature.melspectrogram() function, and then the spectrum can be displayed by using the librosa.display.specshow() function. The spectrograms of all speech segments in the SUSP-SFD corpus are extracted, and the spectrogram image data set is constructed as Figure 12 shows.

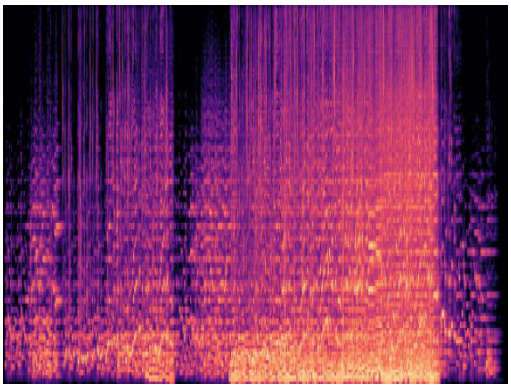

Figure 12. Spectrogram (example)

The spectrogram dataset is named according to the naming method of the corpus, and it is available on: https://github.com/csx0709/spectrogram. In the dataset constructed this time, the fatigue label is defined as 1, the normal label is defined as 0, and the number of images for fatigue is 116 and for normal is 94. Spectrogram image data set is shown in Table 1.

Table 1. Fatigue and normal spectrogram image dataset

Category	Label	Quantity/Piece
Fatigue	1	116
Normal	0	94

4.4 Fatigue Spectrogram Classification

This paper completes the use of the relevant models of the atlas data through the construction of the CNN-ELM model, realizes the evaluation and prediction of the model through the subsequent compilation and model fitting, detects the relevant fatigue information of the data, and obtains the final implementation effect of the system.

The CNN-ELM model we designed has an input layer, 5 convolution layers, 3 pooling layers, 2 standardization layers and ELM classification layers. The activation function we used is Relu. The network parameters of the CNN-ELM hybrid model are shown in Table 2.

Table 2. Network parameters of CNN-ELM hybrid model

Layer_name	Input_size	Convolutional Kernel Length	Channels	Step	Pooled domain size	Output_size
Input			3			227×227
Convo1	227×227	11	96	4		55×55×96
Pool1	55×55×96			2	3×3	27×27×96
Stand1	27×27×96					27×27×96
Convo2	27×27×96	5	256	1		27×27×256
Pool2	27×27×256			2	3×3	13×13×256
Stand2	13×13×256					13×13×256
Convo3	13×13×256	3	384	1		13×13×384
Convo4	13×13×384	3	384	1		13×13×384
Convo5	13×13×384	3	256	1		13×13×256
Pool3	13×13×256			2	3×3	6×6×256
ELM						

4.5 Experimental Results

In this paper, fatigue classification based on CNN-ELM is summarized by drawing confusion matrix, which is shown in Table 3. Among them, the accuracy of fatigue degree identification is 73.8%, the misclassification rate is 26.2%, and the accuracy is 74.29%.

Table 3. Confusion matrix of CNN-ELM experiment

	Normal (Predicted)	Fatigue (Predicted)
Normal (True)	94	32
Fatigue (True)	22	62

4.5.1 Result Analysis for RQ1

Approach. For this research question, we would like to investigate how effectively we can classify fatigue by speech using our proposed approach.

In order to fully validate the performance of the CNN-ELM hybrid model, a comparison was made with traditional algorithms CNN and commonly used algorithms Mask R-CNN, MR R-CNN and Faster R-CNN in terms of spectrogram image recognition accuracy and training time. The comparison results are shown in Table 4.

Results. From Table 4, it can be seen that the recognition accuracy of the two models is very close, with only a difference of 2.86%, but there is a significant difference in training time. The training time of the CNN-ELM is only 34.67% of the CNN. Based on the above analysis, it can be concluded that ELM classifiers have fast training speed and can achieve high-quality training results. In a word, the mixture model can not only quickly and automatically adjust the parameters, extract the discriminant features that are conducive to classification, but also obtain better classifier parameters. The experimental results intuitively reflect the fast advantage of the CNN-ELM hybrid model, solving the problem of long training time for CNN models.

In addition, this article uses the commonly used algorithms Mask R-CNN, MR R-CNN and Faster R-CNN for comparative experiments, and the data is shown in Table 4. Our method higher accuracy and lower training time than Mask R-CNN and Faster R-CNN. In contrast to MR R-CNN, although the accuracy of our method is slightly lower than MR R-CNN, the training time is slightly faster.

In summary, the method we proposed has better overall performance. So using the CNN-ELM hybrid model to recognize fatigue spectrum images has more application prospects.

Table 4. Comparison of recognition accuracy and training time

Model	Accuracy	Training time
CNN [27]	71.43%	675s
Mask R-CNN [28]	72.86%	456s
MR R-CNN [29]	75.71%	473s
Faster R-CNN [30]	72.38%	318s
CNN-ELM (Ours)	74.29%	234s

4.5.2 Result Analysis for RQ2

Approach. For this research question, we would like to combine CNN with other classifiers and compare the recognition accuracy and training time with our proposed method. We compare the classification accuracy of the CNN-ELM hybrid model with the CNN-SVM model [32]. The comparison results are shown in Table 5.

Table 5. Comparison of different classifiers

Model	Accuracy	Training time
CNN [27]	71.43%	675s
CNN-SVM [32]	72.38%	526s
CNN-ELM (Ours)	74.29%	234s

Results. From Table 5, we can find that the recognition accuracy of our proposed approach is higher than CNN-SVM. The training time is less than half that of CNN-SVM. So it's not difficult for us to summarize that, in the application of fatigue detection using speech spectrograms, the combination of CNN and ELM performs better than the combination of classical classifier SVM.

4.5.3 Result Analysis for RQ3

Approach. For this research question, we would like to adopt cross-validation. 80% of the data is randomly sampled using a stratified way as training data, 10% of the data is randomly sampled as verification data, and the remaining 10% is used as the test data set. Firstly, the training set is used to train the classifier, and then the dev set is used to test the trained model as a performance indicator to evaluate the classifier.

Results. We can avoid overfitting and data leakage during the training process.

5 Conclusion

This article innovatively transforms fatigue classification based on speech feature parameters into fatigue image recognition based on spectrograms. Hybrid model combining convolutional neural network and extreme learning machine is applied to fatigue classification based on spectrogram. CNN is used to extract features from the input image. The feature mapping will eventually be encoded into a one-dimensional vector and sent to ELM for classification. The detailed design of the CNN-ELM model is given, including parameter design, structural analysis and the derivation of the back propagation in the iterative process. The CNN-ELM model designed has an input layer, 5 convolution layers, 3 pooling layers, 2 standardization layers and ELM classification layers. The activation function used is Relu.

The CNN-ELM model designed in this article has an accuracy of 74.29% for fatigue classification based on spectrogram image recognition, which is 2.86% higher than the CNN model recognition accuracy. The training time is only 34.67% of the CNN model, which solves the problem of long training time for CNN models. Compared with commonly used model Mask R-CNN, MR R-CNN and Faster R-CNN, our method has more stronger overall

performance. Therefore, the spectral image recognition method of CNN-ELM has more application prospects. Subsequent work will expand the corpus to achieve better accuracy.

For the future work, we will still strive to improve the accuracy of fatigue classification based on speech analysis. We will focus on data mining of spectrogram and feature fusion of primary and secondary networks.

Acknowledgments

This work is sponsored by: (1) the Universities Natural Science Research Projects of Jiangsu Province (No. 22KJB520032); (2) Nantong Institute of Technology's Second Batch of Middle and Young Backbone Teacher Training Special Project (No. ZQNGGJS202225); (3) the Science and Technology Planning Project of Nantong City (No. JCZ21058, No. JCZ2023075, No. MS2024016); (4) the Universities Philosophy and Social Science Research Projects of Jiangsu Province (No. 2023SJYB1720 and 2022SJYB1758); (5) Industry-Academia-Research Collaborative Innovation Foundation of Chinese Universities-Science and Education Project of Digital Intelligence (No. 2023RY001).

References

[1] D. Pawlik, D. Mroczek, Fatigue and Training Load Factors in Volleyball, *International Journal of Environmental Research and Public Health*, Vol. 19, No. 18, Article No. 11149, September, 2022.

[2] S. Ahmed, M. Joyo, H. Mahdi, I. Kiwarkis, Muscle Fatigue Detection and Analysis Using EMG Sensor, *2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Kuala Lumpur, Malaysia, 2020, pp. 1-4.

[3] A. Barsotti, K. Khalaf, D. Gan, Muscle fatigue evaluation with EMG and Acceleration data: a case study, *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 2020, pp. 3138-3141.

[4] K. Heaton, J. Williamson, A. Lammert, K. Finkelstein, C. Haven, D. Sturim, C. Smalt, T. Quatieri, Predicting changes in performance due to cognitive fatigue: A multimodal approach based on speech motor coordination and electrodermal activity, *The Clinical Neuropsychologist*, Vol. 34, No. 6, pp. 1190-1214, July, 2020.

[5] X. Li, G. Li, L. Peng, L. Yan, C. Zhang, Driver fatigue detection based on speech feature transfer learning, *Journal of the China Railway Society*, Vol. 42, No. 4, pp. 74-81, April, 2020.

[6] S. Chen, Y. Sun, H. Zhang, Q. Liu, X. Lv, X. Mei, Speech Fatigue Detection Based on Deep Learning, *Journal of Physics: Conference Series*, Vol. 2224, Article No. 012023, April, 2022.

[7] M. Shahrin, L. Wyse, Applying visual domain style transfer and texture synthesis techniques to audio: insights and challenges, *Neural Computing and Applications*, Vol. 32, No. 4, pp. 1051-1065, February, 2020.

[8] Y. Zhang, S. Dai, W. Song, L. Zhang, D. Li, Exposing speech resampling manipulation by local texture analysis on spectrogram images, *Electronics*, Vol. 9, No. 1, Article No. 23, January, 2020.

[9] V. Gupta, S. Juyal, Y. Hu, Understanding human emotions through speech spectrograms using deep neural network, *The Journal of Supercomputing*, Vol. 78, No. 5, pp. 6944-6973, April, 2022.

[10] H. Lv, *Research and Implementation of Light Weighted Semantic Segmentation Based on Ege Information*, Master's Thesis, Beijing University of Posts and Telecommunications, Beijing, China, 2020.

[11] A. Ivakhnenko, V. Lapa, I. Technica, R. Mcdonough, *Cybernetics and forecasting techniques*, American Elsevier, 1967.

[12] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, *Neural computation*, Vol. 18, No. 7, pp. 1527-1554, July, 2006.

[13] W. Wong, Y. Shi, Y. Qi, R. Golden, Using an RBF neural network to locate program bugs, *2008 19th International Symposium on Software Reliability Engineering (ISSRE)*, Seattle, Washington, USA, 2008, pp. 27-36.

[14] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, *Communications of the ACM*, Vol. 60, No. 6, pp. 84-90, June, 2017.

[15] A. Khan, A. Sohail, U. Zahoora, A. Qureshi, A Survey of the Recent Architectures of Deep Convolutional Neural Networks, *Artificial Intelligence Review*, Vol. 53, No. 8, pp. 5455-5516, December, 2020.

[16] Y. Zhang, Y. Hao, A Survey of SAR Image Target Detection Based on Convolutional Neural Networks, *Remote Sensing*, Vol. 14, No. 24, Article No. 6240, December, 2022.

[17] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, Y. Miao, Review of Image Classification Algorithms Based on Convolutional Neural Networks, *Remote Sensing*, Vol. 13, No. 22, Article No. 4712, November, 2021.

[18] G. Lagani, F. Falchi, C. Gennaro, G. Amato, Comparing the performance of Hebbian against backpropagation learning using convolutional neural networks, *Neural Computing and Applications*, Vol. 34, No. 8, pp. 6503-6519, April, 2022.

[19] W. Wu, Y. Pan, Adaptive Modular Convolutional Neural Network for Image Recognition, *Sensors*, Vol. 22, No. 15, Article No. 5488, August, 2022.

[20] N. Baker, N. Zengeler, U. Handmann, A Transfer Learning Evaluation of Deep Neural Networks for Image Classification, *Machine Learning and Knowledge Extraction*, Vol. 4, No. 1, pp. 22-41, March, 2022.

[21] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 12, pp. 6999-7019, December, 2022.

[22] C. Santos, J. Papa, Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks, *ACM Computing Surveys*. Vol. 54, No. 10s, Article No. 213, September, 2022.

[23] Z. Li, H. Jiang, Q. Mei, Z. Li, Forest Fire Recognition Based on Lightweight Convolutional Neural Network, *Journal of Internet Technology*, Vol. 23, No. 5, pp. 1147-1154, September, 2022.

[24] T. Ahmed, P. Das, M. Ali, M. Mahmud, A Comparative Study on Convolutional Neural Network Based Face Recognition, *2020 11th International Conference on Computing, Communication and Networking Technologies*

(ICCCNT), Kharagpur, India, 2020, pp. 1-5.

[25] J. Bharadiya, Convolutional Neural Networks for Image Classification, *International Journal of Innovative Science and Research Technology*, Vol. 8, No. 5, pp. 673-677, May, 2023.

[26] Q. Yao, J. Cai, J. Qiu, W. Li, J. Qu, Image recognition of blade surface integrity based on VGG-16 Convolutional neural network, *Aviation Precision Manufacturing Technology*, Vol. 57, No. 5, pp. 4-8, October, 2021.

[27] J. Chu, Z. Guo, L. Leng, Object Detection Based on Multi-Layer Convolution Feature Fusion and Online Hard Example Mining, *IEEE Access*, Vol. 6, pp. 19959-19967, March, 2018.

[28] B. Corrigan, Z. Tay, D. Konovessis, Real-Time Instance Segmentation for Detection of Underwater Litter as a Plastic Source, *Journal of Marine Science and Engineering*, Vol. 11, No. 8, Article No. 1532, August, 2023.

[29] Y. Zhang, J. Chu, L. Leng, J. Miao, Mask-Refined R-CNN: A Network for Refining Object Details in Instance Segmentation, *Sensors*, Vol. 20, No. 4, Article No. 1010, February, 2020.

[30] M. Sahin, H. Ulutas, E. Yuce, M. Erkoc, Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images, *Neural Computing and Applications*, Vol. 35, No. 18, pp. 13597-13611, June, 2023.

[31] S. Chen, *Research of fatigue detection based on speech analysis*, Master's Thesis, Soochow University, Suzhou, China, 2017.

[32] K. Prema, J. Visumathi, Hybrid Approach of CNN and SVM for Shrimp Freshness Diagnosis in Aquaculture Monitoring System using IoT based Learning Support System, *Journal of Internet Technology*, Vol. 23, No. 4, pp. 801-810, July, 2022.

Biographies

Shuxi Chen received her M.S. degree in Information and Communication Engineering from Soochow University in 2017. She is a lecturer of Computer and Information Engineering School in Nantong Institute of Technology. Her recent research interests are machine learning and speech recognition.

Yiyang Sun received his M.S. degree in Electronic and Communication Engineering from Hohai University in 2016. He is a lecturer of Computer and Information Engineering School in Nantong Institute of Technology. His recent research interests are machine learning and deep learning.

Jianlin Qiu received his M.S. degree in Electrical Engineering from Shanghai University in 2005. He is a professor of Information Science and Technology School in Nantong University. His recent research interests are machine learning, databases, GIS, semantic web and web intelligence.

Haifei Zhang received his M.S. degree in Software Engineering from Soochow University in 2005. He is a professor of Computer and Information Engineering School of Nantong Institute of Technology. His recent research interests are machine learning, databases and GIS.

Hao Chen received his Ph.D. degree in 2022 from School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden. His recent research interests are distributed machine learning, large models and edge computing.